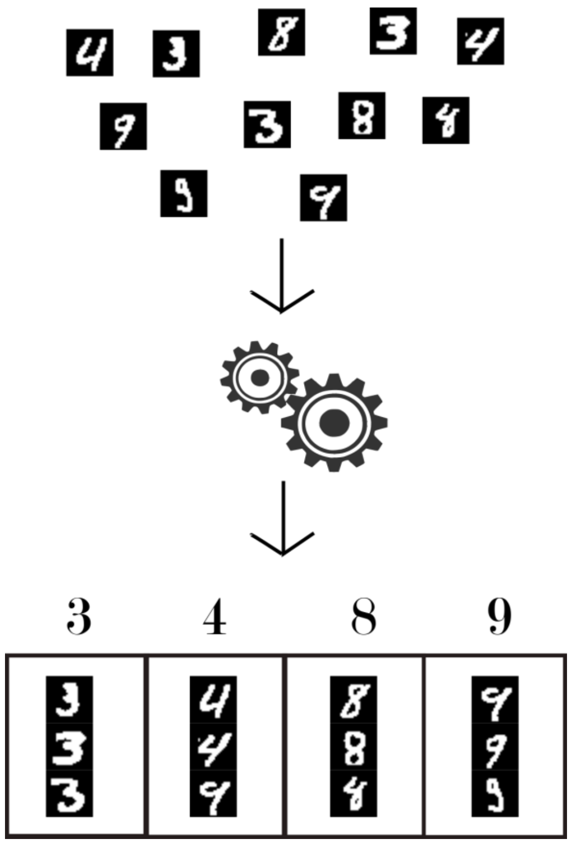
RANDOM FOREST-BASED DIFFUSION INFORMATION GEOMETRY FOR SUPERVISED VISUALIZATION AND DATA EXPLORATION

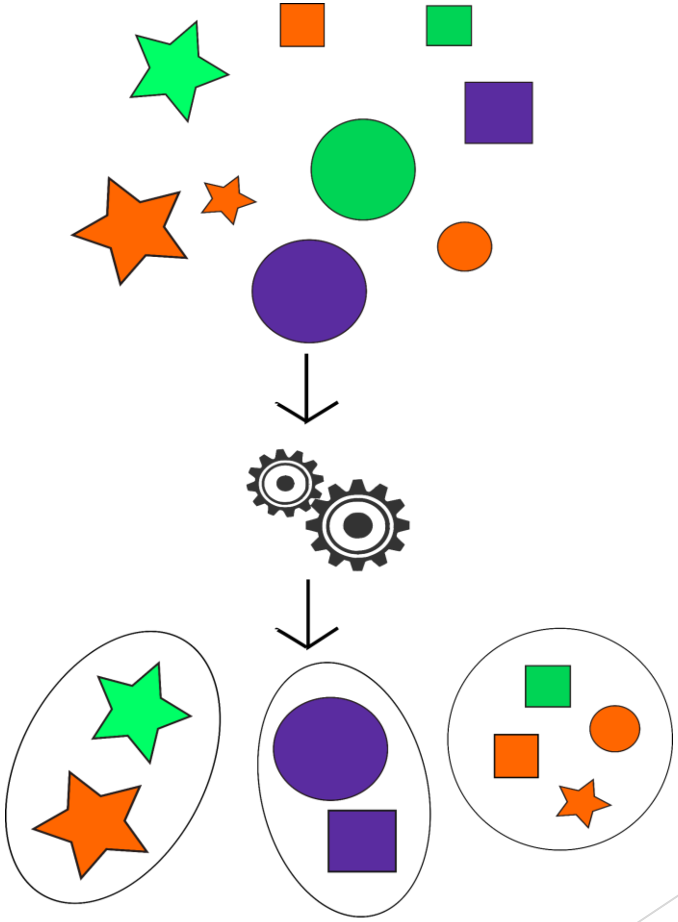
Jake S. Rhodes, Adele Cutler, Guy Wolf, Kevin R. Moon

Machine Learning

Supervised (Labeled Data)



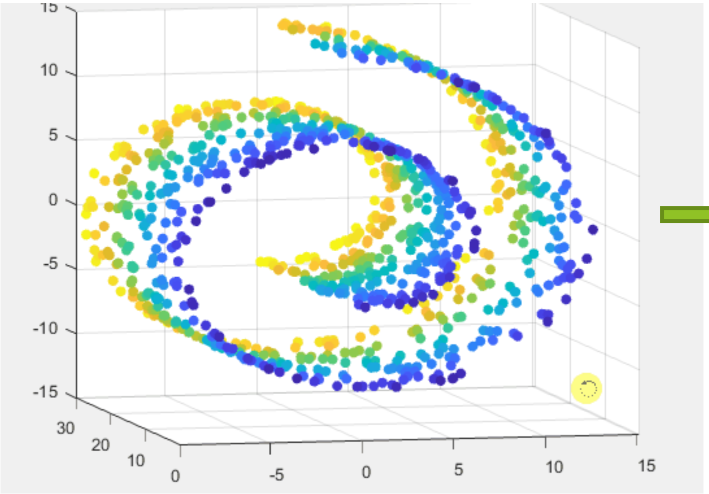
Unsupervised (Unlabeled)



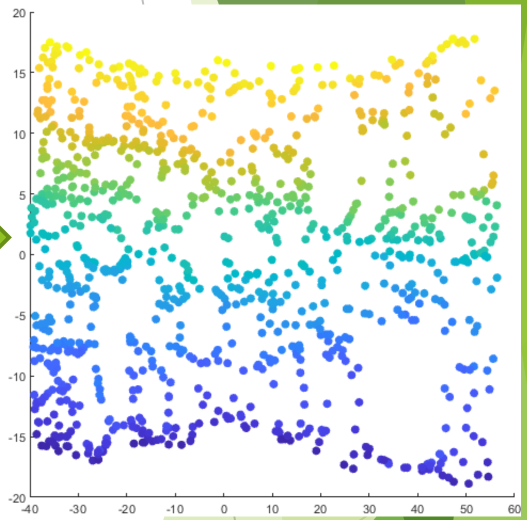
Dimensionality Reduction (Unsupervised)

- Decrease data complexity
- Preprocessing and Visualization
- Three major types [5]:
  - Principal Components (linear)[8, 10, 16, 21]
  - Matrix Factorization (linear) [9, 11]
  - **Manifold Learning (non-linear)** [6, 13, 14, 15, 18, 19]

3D Swiss Roll



2D Embedding



How can we incorporate extra information (e.g. class labels) into dimensionality reduction?

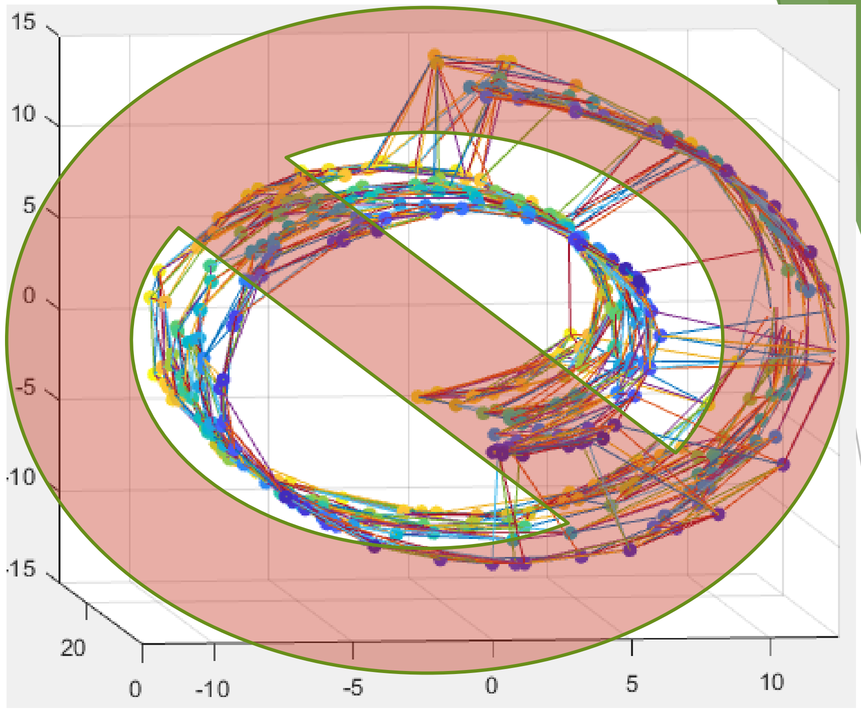
Manifold Learning

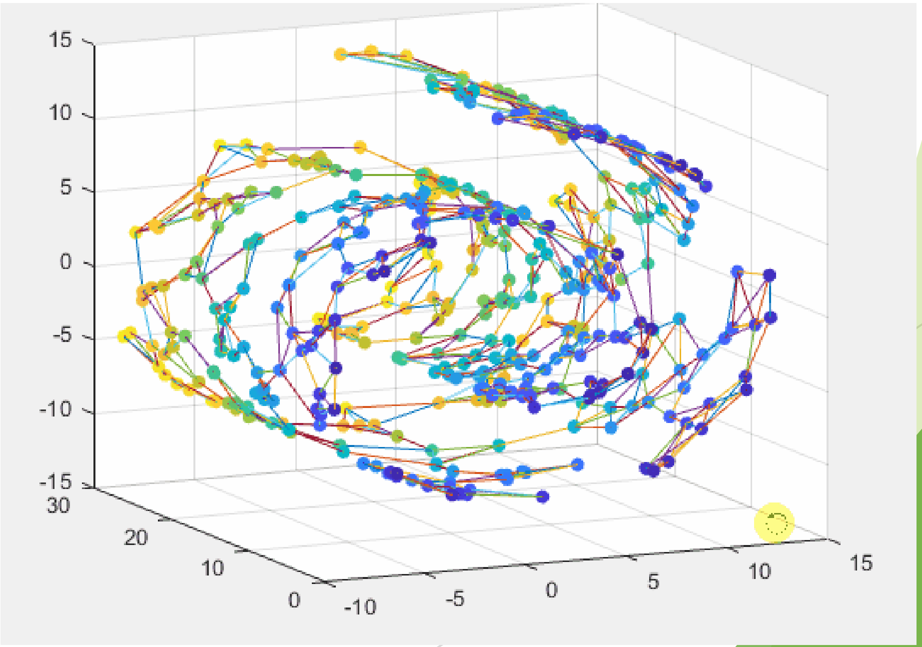
Starting Local:

- Calculate Pairwise Distances
- Keep only k nearest points
- Small distances to keep from “exiting” the manifold

Moving Global:

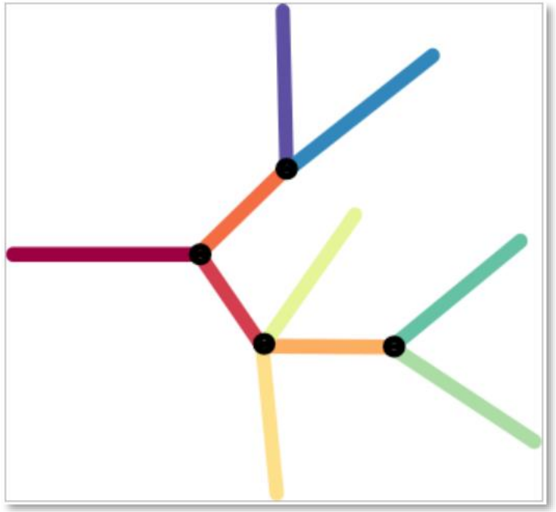
- “Walking” the edges (Diffusion [6, 15])

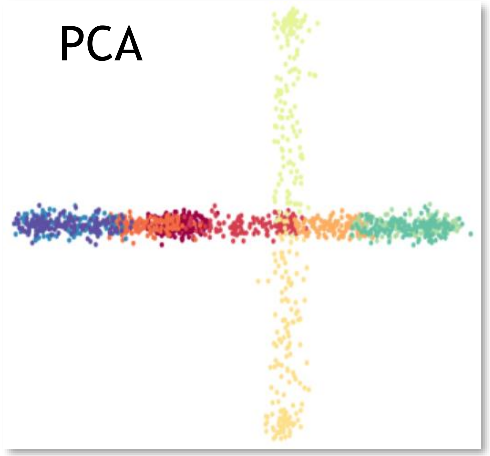




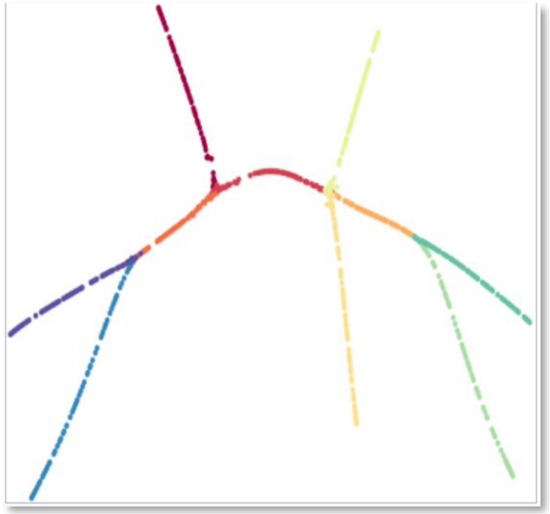
Manifold Learning (PHATE [14], 2019)

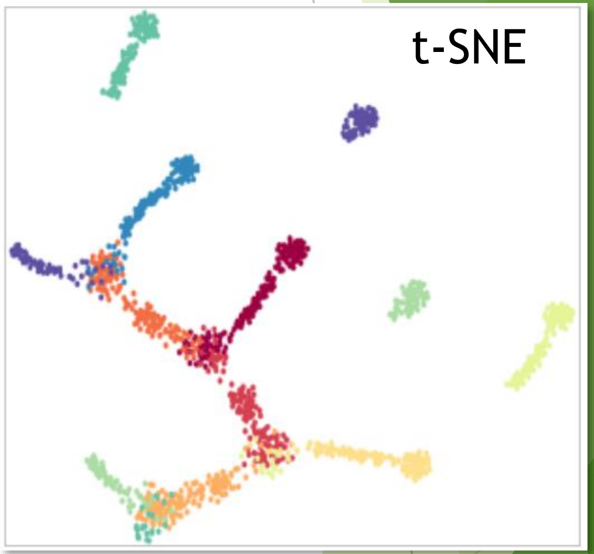
Artificial Tree Data (60 dimensions)



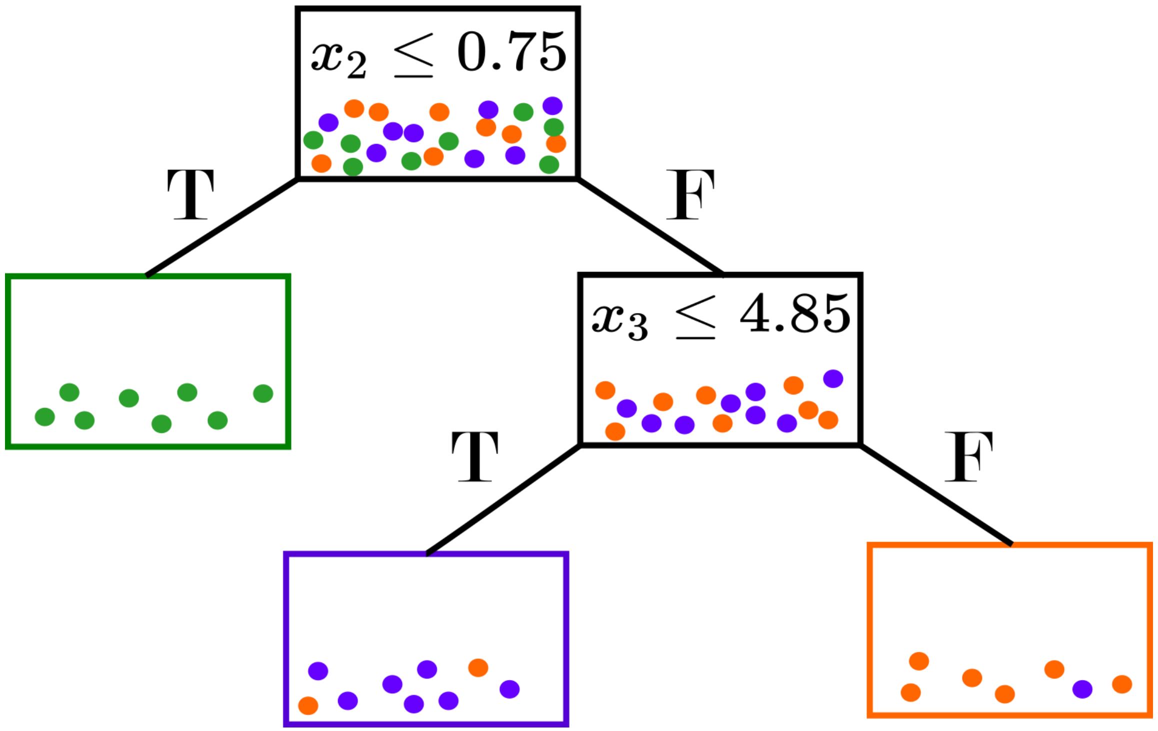


PHATE



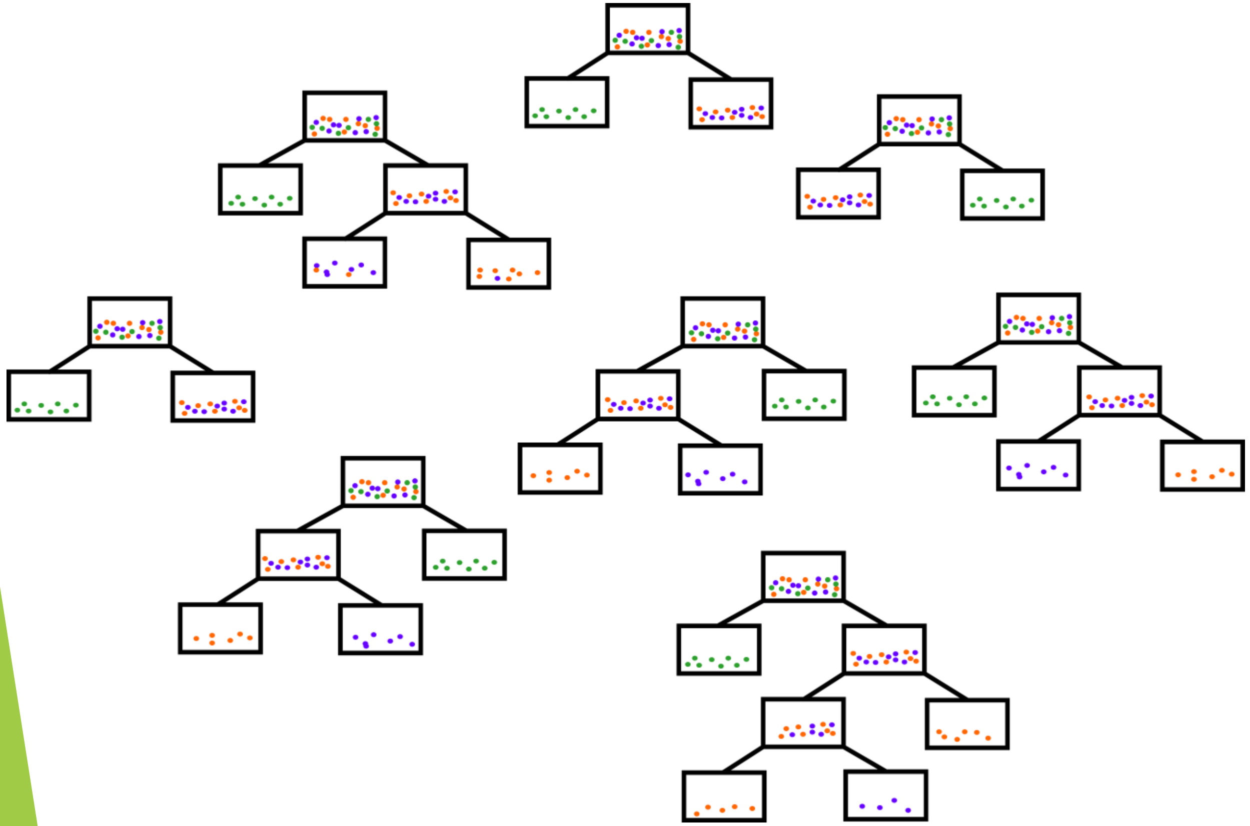


Decision Tree Classification (Supervised) [4]



- Binary Variable Splits
- Majority-Vote Classification
- Terms:
  - Root Node
  - Splitting Node
  - Leaf Node

Random Forests (Supervised) [2]

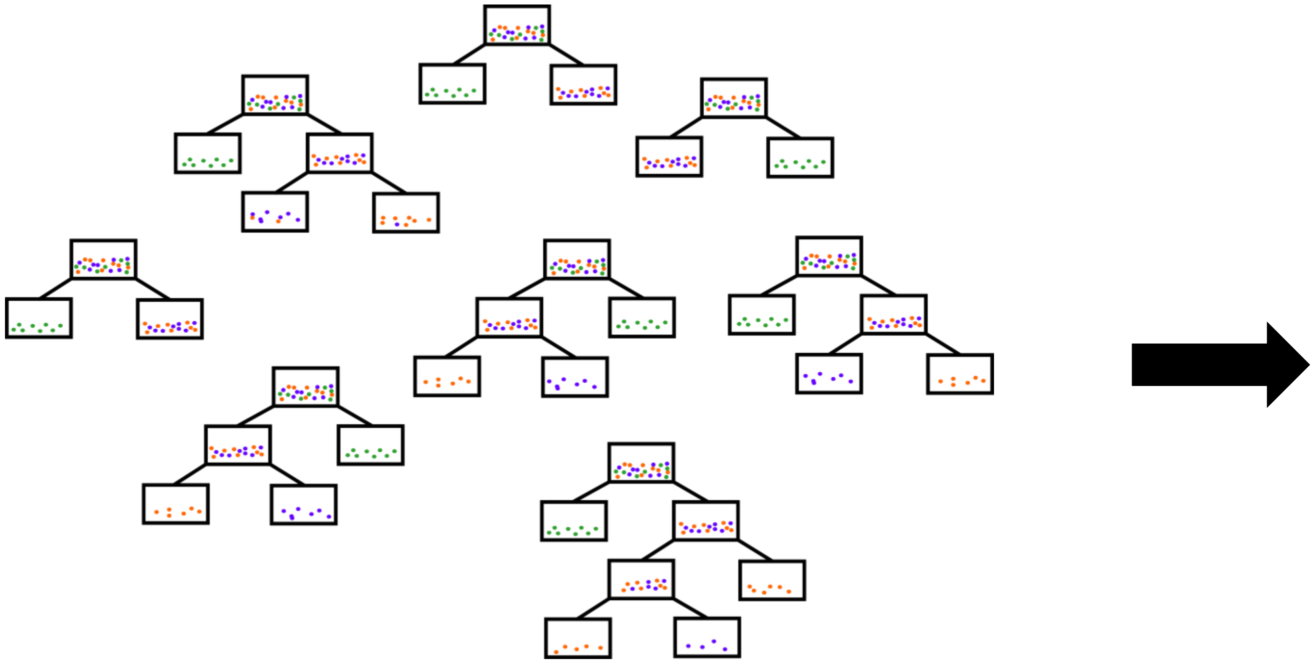


- Ensemble of Decision Trees
- Two-Part Randomization
  - Bootstrap Sampling
  - Random Variable-Splitting Selection

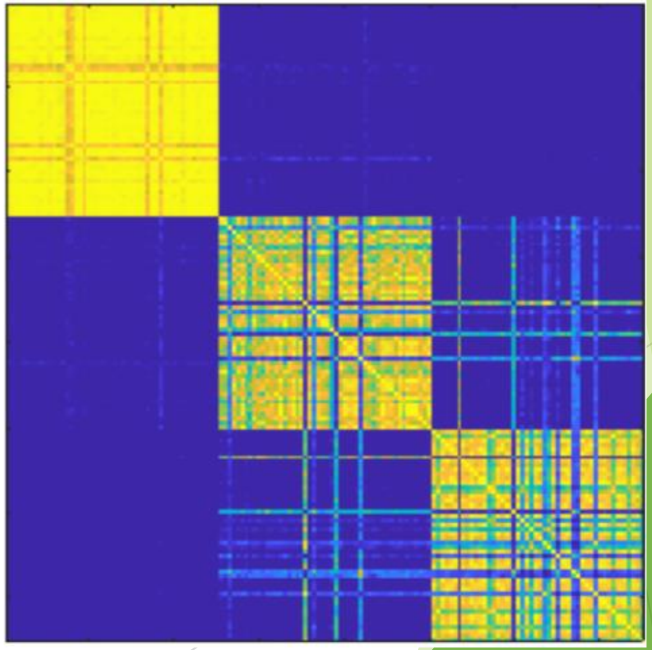
Random Forest Proximities [2, 3]

The proximity between two observations is the proportion of trees in which they reside in the same terminal node.

- Adaptive similarities [12]
- Proximities capture variable importance [2, 12]
- Idea: Use proximities as kernel for PHATE [14]

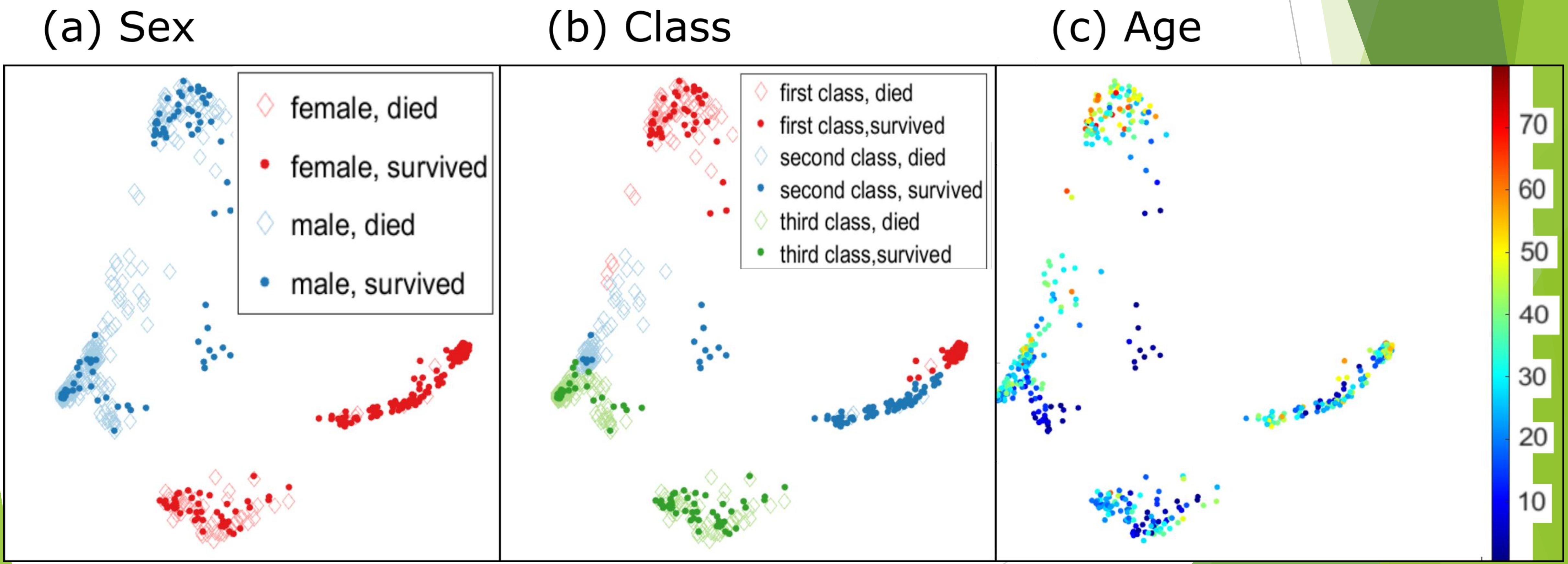


Proximity (Affinity) Matrix



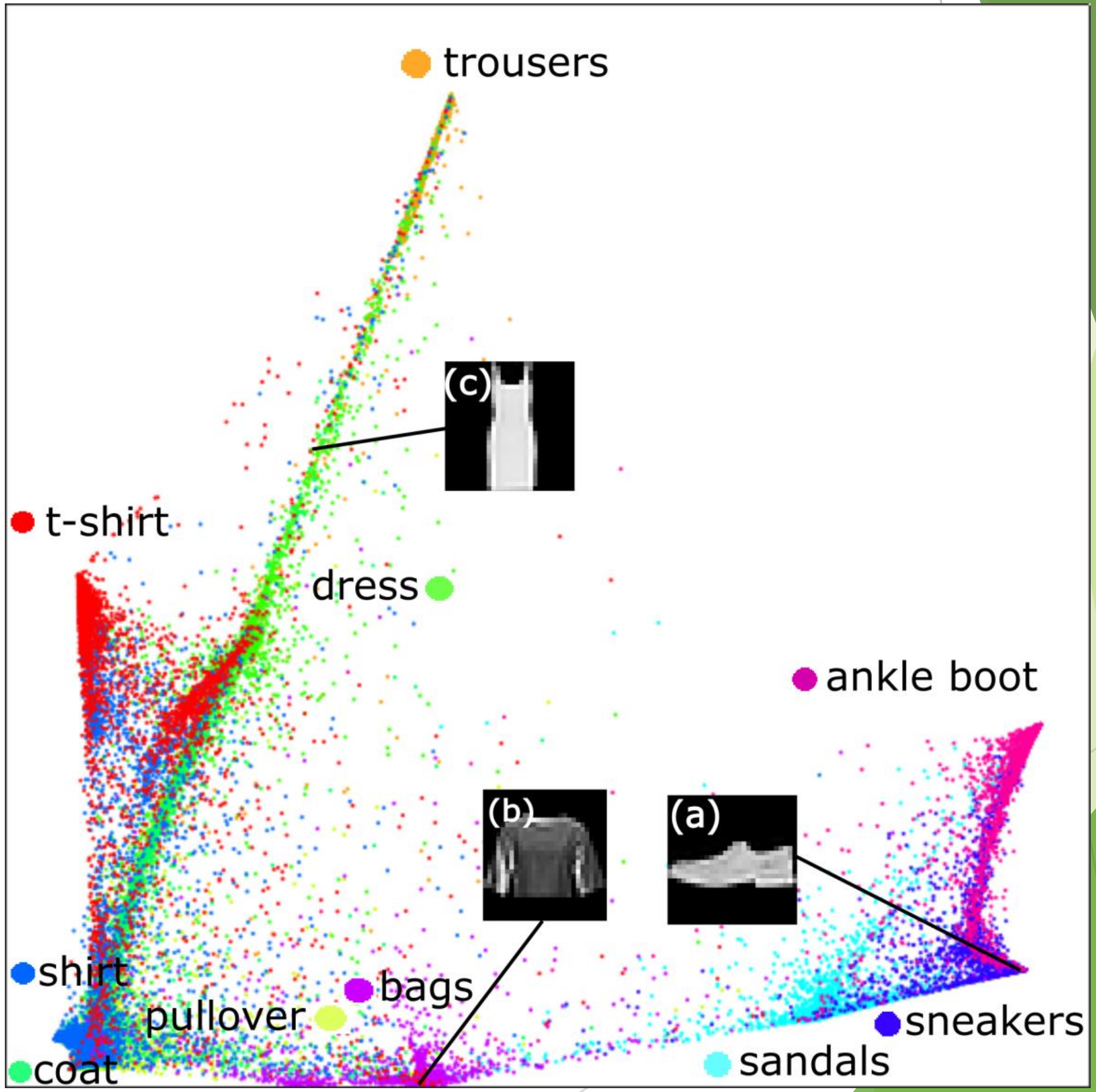
Example: Titanic [7]

- 2 class labels (died or survived)
- 891 observations
- 12 variables

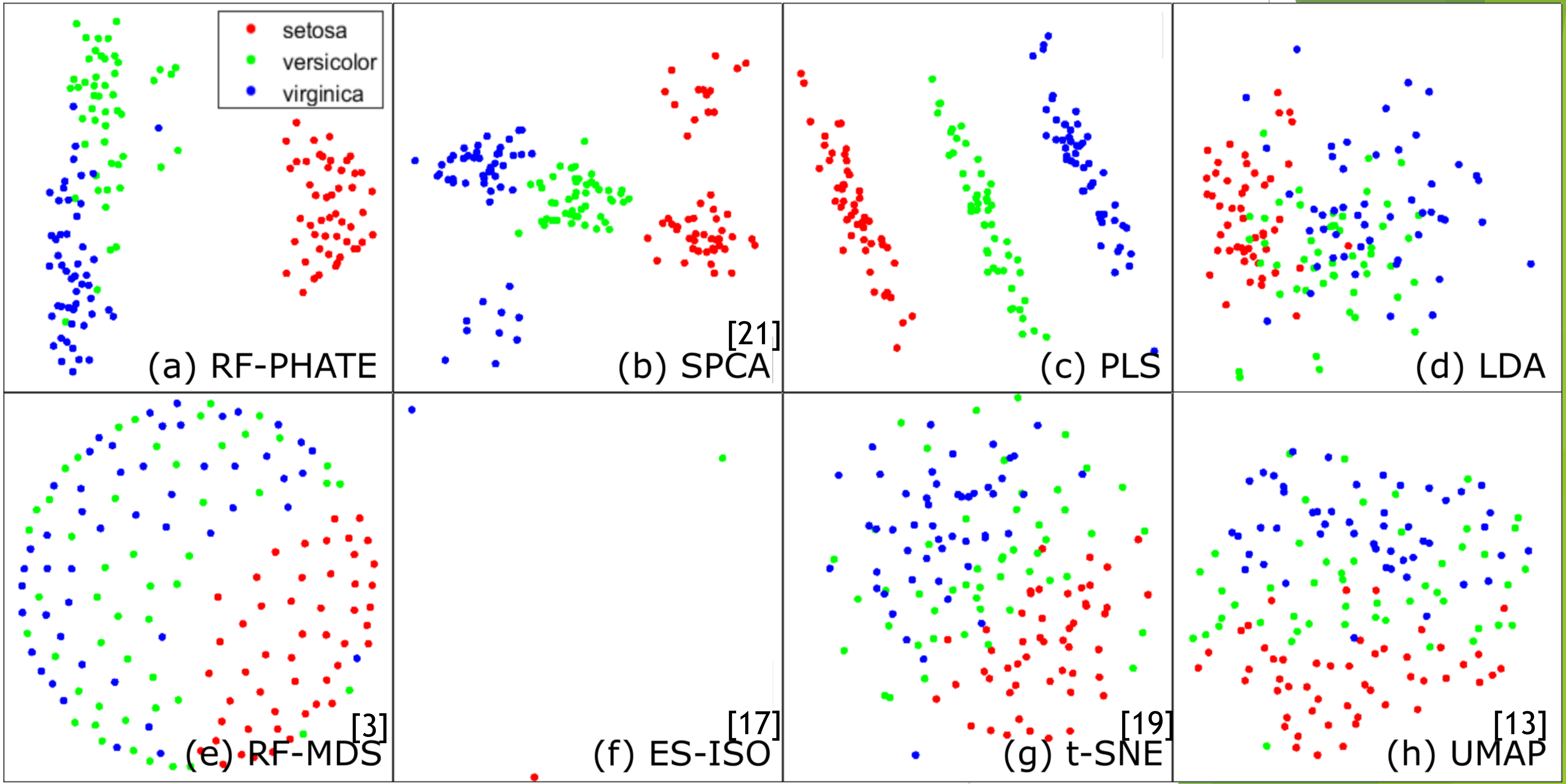


Example: Fashion MNIST [20]

- 10 class labels
- 60,000 images
- 28 x 28 pixels (784 variables)

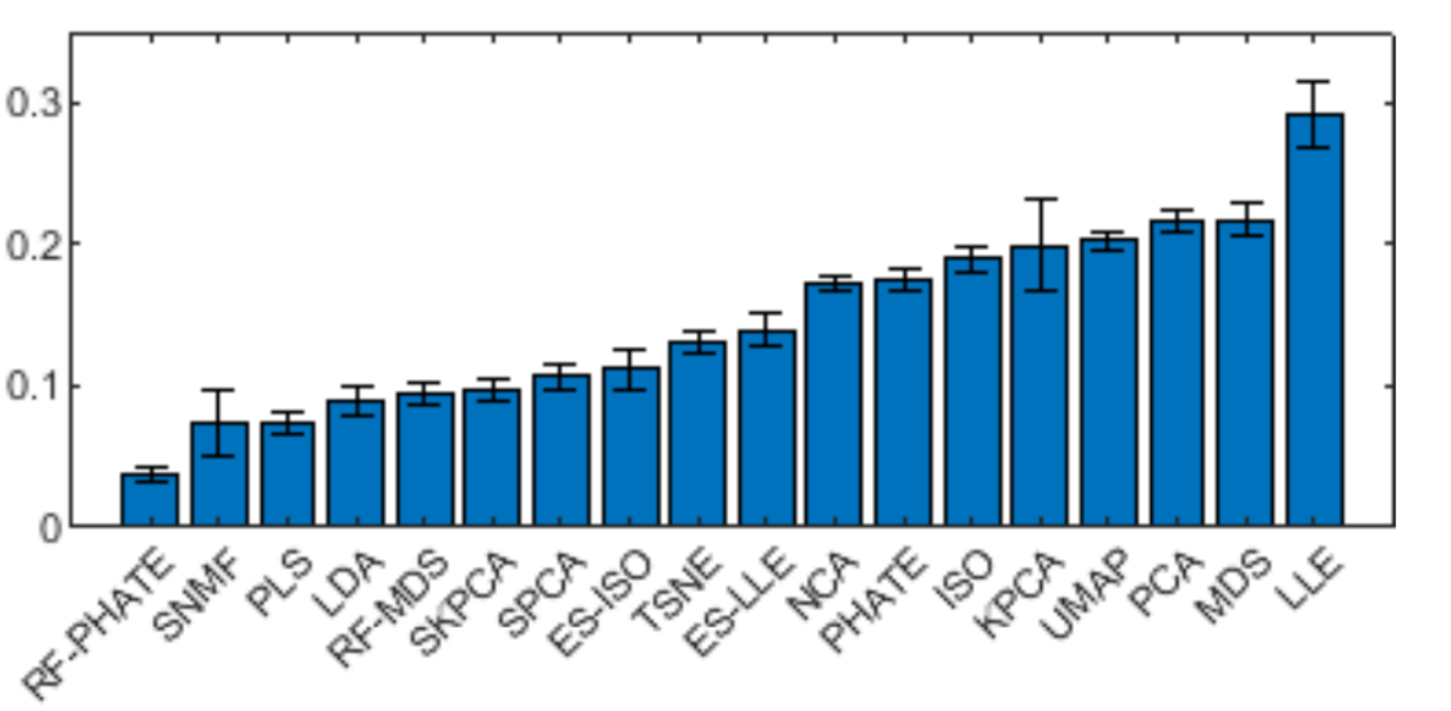


Iris [1] with 1000 Noise Variables



Quantitative Evaluation – Capturing Variable Importance

- Low-dimensional embedding should capture variable importance
- Assess variable importance on original and low-dimensional data
- Compute the correlation between importance measures
- Standardize across datasets (lower is better)



Future:

- New proximity definition to better capture data geometry
- Use as regularization in neural network (autoencoder)
- Incorporate unlabeled version

References I

[1] E. Anderson. The species problem in Iris. *Ann. Missouri Bot*, 23(3):457–509, 1936.

[2] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[3] L. Breiman and A. Cutler. Random forests for scientific discovery. `http://www.math.usu.edu/adele/RandomForests/ENAR.pdf`, (Accessed on 04/15/2020).

[4] L. Breiman, J. Friedman, R. Olshen, et al. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984. new edition (4)?

[5] G. Chao, Y. Luo, and W. Ding. Recent advances in supervised dimension reduction: A survey. *Mach. Learn. Knowl. Extr.*, 1(1):341–358, Jan 2019.

[6] R. R. Coifman and S. Lafon. Diffusion maps. *Appl Comput Harmon Anal*, 21(1):5 – 30, 2006.

[7] P. Hendricks. *titanic: Titanic Passenger Survival Data Set*, 2015. R package version 0.1.0.

[8] H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol*, 1933.

[9] Y. Jia, S. Kwong, J. Hou, et al. Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization. *IEEE Trans Neural Netw Learn Syst*, pages 1–12, 2019.

[10] J.B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[11] D.D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999.

[12] Y Lin and Y Jeon. Random forests and adaptive nearest neighbors. *JASA*, 101(474):578–590, 2006.

References II

[13] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, abs/1802.03426, 2018.

[14] K. R. Moon, D. van Dijk, Z. Wang, et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol*, 37(12):1482–1492, Dec 2019.

[15] B. Nadler, S. Lafon, I. Kevrekidis, et al. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In *NeurIPS*, pages 955–962, 2006.

[16] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *London Edinburgh Philos. Mag. J. Sci*, 2(11):559–572, 1901.

[17] B. Ribeiro, A. Vieira, and J. Carvalho das Neves. Supervised isomap with dissimilarity measures in embedding learning. In *Progress in Pattern Recognition, Image Analysis and Applications*, pages 389–396. Springer Berlin Heidelberg, 2008.

[18] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[19] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *J Mach Learn Res*, 9(Nov):2579–2605, 2008.

[20] H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, abs/1708.07747, 2017.

[21] S. Yu, K. Yu, V. Tresp, et al. Supervised probabilistic principal component analysis. In *KDD*, page 464–473, 2006.